Artificial Intelligence Threat Reporting and Incident Response System

D2.4 Human factors for co-design methodology

Project Title:	Artificial Intelligence Threat Reporting and Incident Response System
Project Acronym:	IRIS
Deliverable Identifier:	D2.4
Deliverable Due Date:	31/12/2022
Deliverable Submission Date:	28/12/2022
Deliverable Version:	V1.0
Main author(s) and Organisation:	Valeria Cesaroni (CEL), Federico Pierucci (CEL), Emanuela Tangari (CEL)
Work Package:	WP2 - System co-design
Task:	Human factors for co-design of effective cross-border threat intelligence sharing
Dissemination Level:	PU: Public

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no 101021727. Content reflects only the authors' view and European Commission is not responsible for any use that may be made of the information it contains.

Quality Control

	Name	Organisation	Date
Editor	Federico Pierucci, Valeria Cesaroni, Emanuela Tangari	CEL	06/12/2022
Peer Review 1	Nikos Kapsalis	KEMEA	02/12/2022
Peer Review 2	Bruno Vidalenc	THALES	14/12/2022
Submitted by (Project Coordinator)	Gonçalo Cadete	INOV	28/12/2022

Contributors

Organisation
CEL
KEMEA
THALES
INOV

Document History

Version	Date	Modification	Partner
0.1	19/10/2022	ToC and Introduction	CEL
0.2	25/11/2022	First version	CEL
0.3	02/12/2022	Internal peer review	KEMEA
0.4	07/12/2022	Peer reviewed by KEMEA version	CEL
0.5	20/12/2022	Peer reviewed by THALES version	CEL
1.0	28/12/2022	Final editing	INOV

Legal Disclaimer

IRIS is an EU project funded by the Horizon 2020 research and innovation programme under grant agreement No 101021727. The information and views set out in this deliverable are those of the author(s) and do not necessarily reflect the official opinion of the European Union. The information in this document is provided "as is", and no guarantee or warranty is given that the information is fit for any specific purpose. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein. The IRIS Consortium members shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials subject to any liability which is mandatory due to applicable law.

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

Abbreviation/ Acronym	Meaning
AI	Artificial Intelligence
ATOS	ATOS IT SOLUTIONS AND SERVICES IBERIA SL
CISCO SPAIN	CISCO SYSTEMS SPAIN S.L.
CEA	COMMISSARIAT A L ENERGIE ATOMIQUE ET AUX ENERGIES ALTERNATIVES
CEL	CYBERETHICSLAB Srls
CERT-RO	CYBERLENS BV
CERTH	ETHNIKO KENTRO EREVNAS KAI TECHNOLOGIKIS ANAPTYXIS
DOI	Diffusion of Innovation
IOR	Inter-organisational Relationship Theory
ECSO	EUROPEAN CYBER SECURITY ORGANISATION
DPIA	Data Protection Impact Assessment
FVH	FORUM VIRIUM HELSINKI OY
GDPR	General Data Protection Regulation
ICCS	INSTITUTE OF COMMUNICATION AND COMPUTER SYSTEMS
INOV	INOV INSTITUTO DE ENGENHARIA DE SISTEMAS E COMPUTADORES, INOVACAO
IMI BCN	INSTITUT MUNICIPAL D'INFORMATICA DE BARCELONA
INTRA	INTRASOFT INTERNATIONAL SA
IoT	Internet of Things
KEMEA	KENTRO MELETON ASFALEIAS
SAT	Social Acceptance of Technology
SID	SIDROCO HOLDINGS LIMITED
TalTech	TALLINNA TEHNIKAÜLIKOOL
TAM	Technology Acceptance Model
THALES	THALES SIX GTS FRANCE SAS
TOE	Technology-Organization-Environment
TRA	Theory of Reasoned Action
TU Delft	TECHNISCHE UNIVERSITEIT DELFT
UPC	UNIVERSITAT POLITECNICA DE CATALUNYA
UTAUT	Unified Theory of Acceptance and Use of Technology

Executive Summary

This deliverable, an outcome of task T2.4 "Human factors for co-design of effective cross-border threat intelligence sharing", aims at providing the **human factors** that will be taken into account in the process of co-designing the IRIS technology. Those factors will be of the utmost importance for granting the multidisciplinary and co-creative development of the IRIS technology, especially considering the IRIS collaborative-first approach. This document has a twofold objective: on the one hand, its empirical goal is to **produce a tailored methodology** based on a specific methodology called Social Acceptance of Technology (SAT), developed by the project partner CEL that will allow for assessing IRIS practitioner's acceptance and engagement to the IRIS technology. On the other hand, it will offer a novel theoretical framework in which it will be possible to **address the social, cultural, and political dimensions** that will allow an in-depth (and beyond state-of-the-art) understanding of the motivating elements toward the future adoption of the IRIS –and similar– technology, aimed at protecting IoT and AI-enabled systems from cyber threats and attacks, by fostering information sharing practices.

With the aim of selecting the most relevant human factors to be considered, the methodology will firstly carry out a literature review taking into account, *inter alia*, the *Ethics Guidelines for Trustworthy AI* (AI HLEG, 2019), as well as the ethical requirements already defined in deliverable D2.3 "*Ethics and data protection requirements specification*". This methodology will then define quantitative and qualitative techniques to be applied within the context of deliverable D2.7 "*IRIS evaluation and impact assessment*" to **assess the social acceptance of IRIS technology** by the main stakeholders (e.g., security practitioners, security services providers, decision makers, etc.). The methodology will define a framework eventually customisable also to other stakeholders other than the IRIS' practitioners, such as for instance the project pilots.

Despite the current methodology has been validated by the project Ethics Board, this is to be considered as an open methodology, that will be constantly updated in the execution of the IRIS technology assessment. The final version will be reported in D2.7 along with the IRIS technology assessment outcome.

1 INTRODUCTION

1.1 Deliverable Purpose

Within the context of project task T2.4, this deliverable aims to describe the IRIS methodology for human factors identification to be used in co-design of effective cross-border threat intelligence sharing.
This achievement is carried out through the definition of a methodology including a model of observation, understanding, and evaluation.
The methodology, along with the model, will define the qualitative and quantitative empirical study techniques, target stakeholder groups, measurement tools (e.g., interviews, survey questionnaires, focus groups) and the related process. The defined methodology will represent a theoretical and empirical foundation upon whom it will be possible to assess the social acceptance of IRIS technology by a comprehensive network of stakeholders, e.g., security practitioners, security services providers, decision makers, etc., established within the context of task T2.6. The methodology will be thus applied in task T2.6 and refined with the support of the involved partners and stakeholders.

1.2 Relation to other project activities

As previously mentioned, the present document is strongly related to the activities that will be carried out in Task T2.6, given that we present the methodology that will be used in the deliverable D2.7 "IRIS evaluation and impact assessment".

On top of that, it also connects with deliverable D2.3 "Ethics and data protection requirements specification", where the ethics requirements that the project needs to follow in the development phase were defined.

Table 1: Relation to other project documents

#ID	Deliverable name	Deliverable description	Submission date
D2.3	Ethics and data protection requirement specification	It aims to specify the requirements related to ethics, data protection and secure sharing of data.	Month 8
D2.7	IRIS Evaluation and Impact Assessment	It provides results from the IRIS social acceptance assessment supported by the network of stakeholders.	Month 24

1.3 Document structure

Section #	Section title	Brief summary
1	Introduction	Provides a brief explanation of the aim of the present deliverable and its structure.
2	State of the Art	Provides a brief overview of the most important frameworks, theories and models in the literature used for understanding human factors and for assessing the social acceptance of technology. Also, it presents some open challenges that remain unaddressed in the literature, and how they are overcome through the so-called Social Acceptance of Technology (SAT) methodology, the IRIS methodology is on top of.
3	IRIS Social Acceptance of Technology	Describes the IRIS methodology to be used to assess the acceptance of the IRIS platform by practitioners and other stakeholders. It dentifies the human factor domains and the related most relevant barriers that hinder acceptance. Then, it describes the assessment process that will be followed as well as the tools and techniques that will be adopted to conduct the assessment and to produce the results.
4	Conclusions	Provides our final remarks, as well as defines the future activities that will be conducted in order to perform the aforementioned assessment.

Table 2: Document structure

2 STATE OF THE ART

Given the increasingly interconnected nature of the economic, social and organisational activities of human societies in the digital world, the interest in how to strengthen cybersecurity shield against cyber threats and attacks through information sharing for organisations and government agencies has grown enormously in recent years.

However, as it is evident from the literature (Boyce & et al., 2011), nowadays there is still a predominantly technical focus on the creation of secure digital spaces, and this is unsatisfactory not only from a theoretical point of view, but also and above all from an operational and practical one (Corradini, 2020).

Therefore, there is an increasing awareness that information security is not a concept that can be investigated from an exclusively technical or organisational point of view (Jeong & et al., 2019). Indeed, building and consolidating efficient and effective systems to protect the security of information and the sharing of data managed and processed by different types of organisations requires a holistic and complex view that is able to take into account various aspects, including technical, organisational, cultural, and social aspects.

With the aim of identifying the most suitable methodology to be applied for the IRIS definition of human factors and technology assessment, the following subsections will provide a brief overview of the most important frameworks in the literature for human factors evaluation (subsection 2.1), as well as methodologies for technology acceptance (subsection 2.2), pointing out how to go beyond the state of the art (subsection 2.3).

2.1 Human factors

With reference to the conceptual frameworks used and developed so far to investigate non-technical factors that influence organisations to adopt or participate in cyber intelligence sharing with their peers, **Kolini's conceptual framework** (Kolini, 2017) seeks to harmonise and go beyond the currently most diriment frameworks in the information sharing and cybersecurity landscape (DOI model – see section 2.2.5, IOR model).

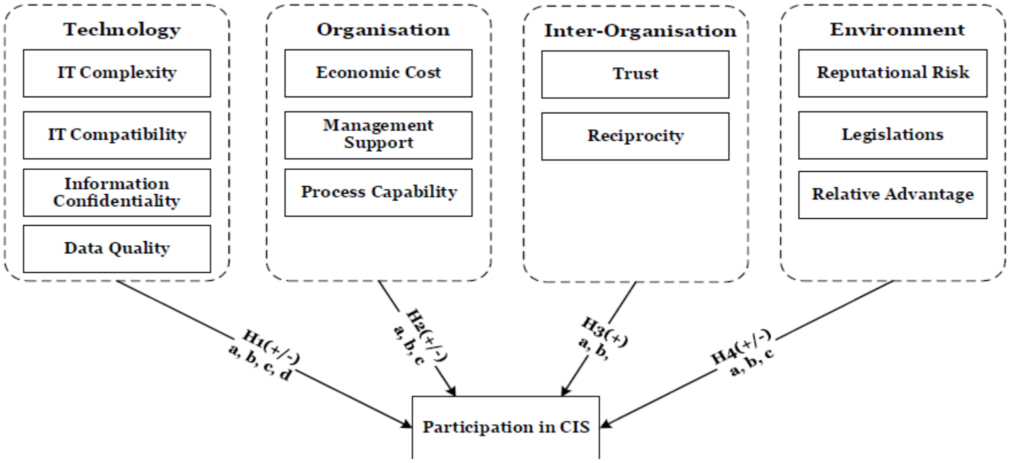


Figure 1: Research Framework from Kolini 2017

The aim of this framework is to draw on the Technology-Organization-Environment (TOE) framework (Tornatzky & Fleischer, 1990) to identify the importance, alongside technological factors, of organisational, intra-organisational and environmental factors in information sharing practices (see Figure 1).

This study is relevant as it represents one of the very first attempts to explore cybersecurity intelligence sharing from an organisational perspective.

Furthermore, it attempts to focus on factors that may impede organisational participation in cybersecurity information sharing activities.

However, **this framework**, while advancing instances of social and organisational understanding of information sharing, **is nevertheless focused on the organisational dimension of information sharing.** Moreover, it does not consider psychological, social and value factors and, more importantly, the intersection between these factors as a determinant in understanding human factors in information sharing as well as the acceptance and appropriateness of certain technologies.

Another extremely valuable framework for the evaluation of artifacts (e.g., models, methods, constructs, instantiations and design theories) in design science research (DSR) is the model developed by (Prat & et al., 2014), which is subdivided into a series of conceptual constructs further articulated into sub-criteria (see Figure 2) ranging from the purely technical and operational to the more social and concerning psychological, cultural, and ethical factors.

Although this framework takes into account a number of items relevant to the understanding of human factors for information sharing, **it bases its analysis on a technical-specific evaluation of the artefact**, which is investigated as an entity on its own and independent of the subjective perception of the artifact.

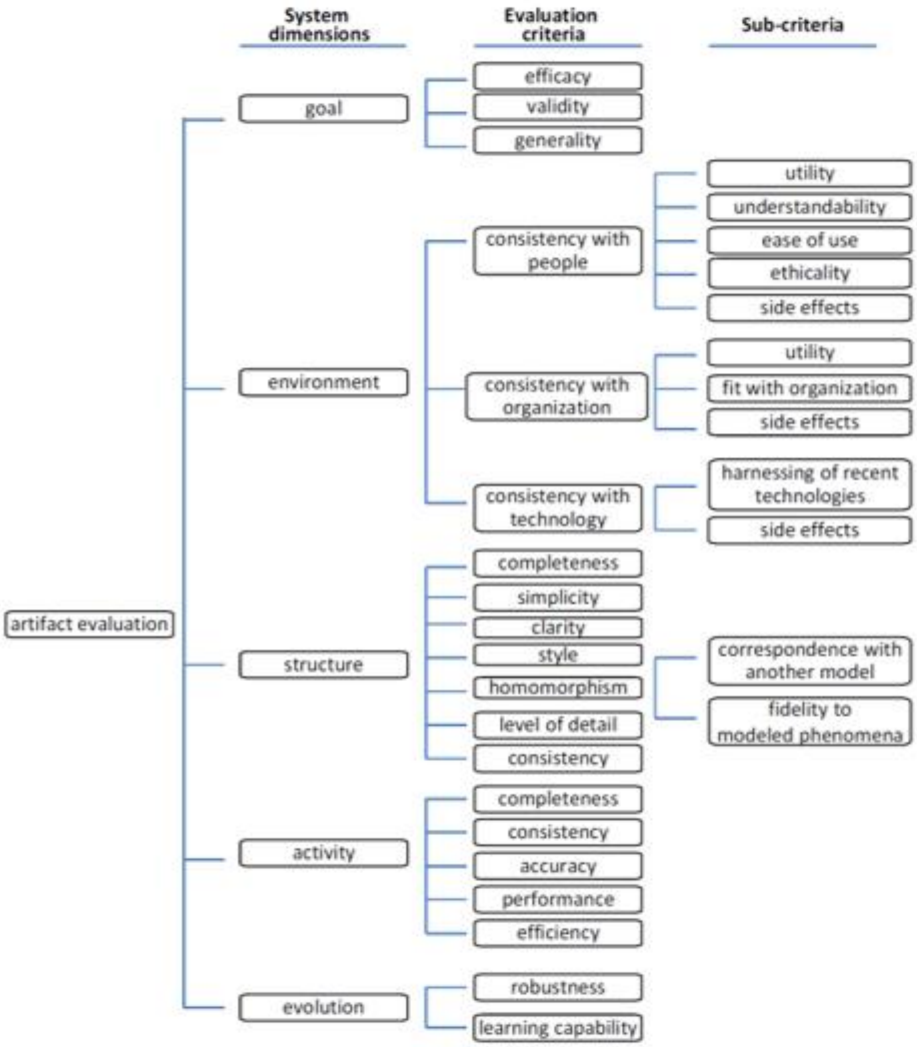


Figure 2: Information System Artifact Evaluation

2.2 Technology acceptance

Despite technological advancements being fundamental passages in humans' history, the problem of **measuring the social acceptance of technology** is relatively new. The reason why this issue was raised only in the last 40 years is that the incredibly rapid advances of information technologies (IT) has boosted innovation in workplaces. Before the IT revolution, in both, workplaces and private life, it was unusual that technological improvements would radically change behaviours and individuals' habits.

The systemic impact of technology on individuals' lives can be understood under the concept of "socio-technical systems", a notion that highlights the interplay between the technological object and the socio-cultural practices that gives meaning to it. In turn, technologies shape our world, and the socio-cultural practices themselves (Ropohl, 1999).

Therefore, the pervasiveness of technologies and the deep impact of IT development on our societies, at every level, have raised more than ever, the issue of measuring social acceptance. Apart from user-technology interaction, the acceptance is also deeply influenced by socio-cultural values implicated, by political and regulatory aspects, historical reasons, needing the contributions of philosophy, value theory, sociology, legal and political experts.

Notwithstanding the complexity of this task, some scholars contributed to the development of a theory of acceptance.

Among them, the most important theories and methods of individual acceptance - therefore not considering in first instance socio-technical implications - were summarised by (Venkatesh & et al., Unified Theory of Acceptance and Use of Technology: A Synthesis and the Road Ahead, 2016); (Kim & Crowston, 2011); (Oliveira & Martins, 2011).

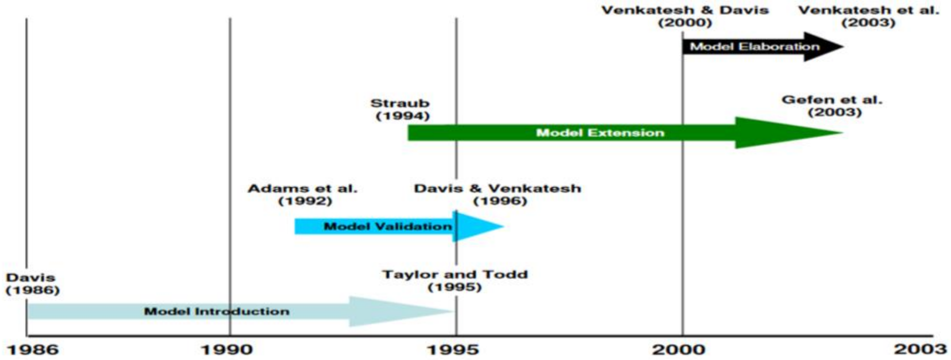


Figure 3: Timeline of technology acceptance theories and models

For the sake of completeness, the most important theories and methods will be illustrated in the next subsections.

2.2.1 Theory of reasoned action

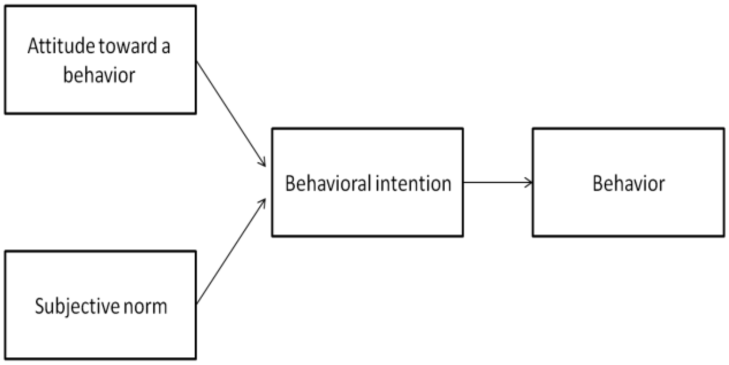


Figure 4: Theory of reasoned action

The Theory of Reasoned Action (TRA) (see Figure 4) was developed by (Fishbein & Ajzen, 1975).

Subjective norm is the person's perception that most people who are important to him think he should or should not perform the behaviour in question. According to (Ramayah & Jantan, 2004) the subjective norms reflect the person's perception of social pressures put on him/her to perform or not to perform the behaviour in question. Subjective norms are a function of normative beliefs. In other words, a person who believes that most people with whom he/she is motivated to comply think he/she should perform the behaviour will perceive social pressure to do so.

2.2.2 Technology Acceptance Model (TAM)

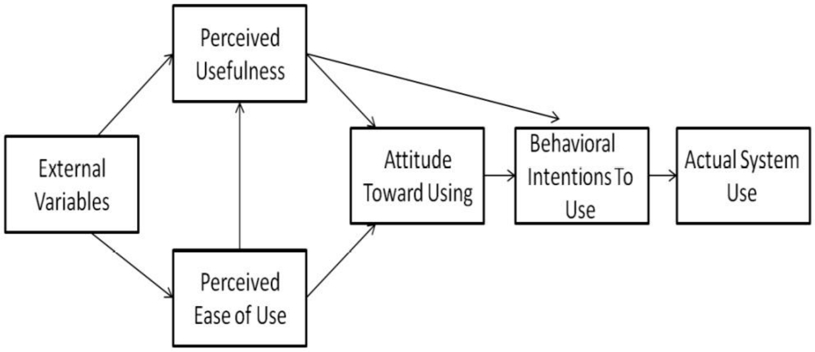


Figure 5: Technology acceptance model

The TAM (see Figure 5) is one of the most widely used methodologies to assess the social acceptance of technologies, and it is mainly focused on workplace environments. The most important dimensions that TAM takes into consideration are perceived usefulness and perceived ease of use. The first is defined by (Davis, 1989) "the degree to which a person believes that using a particular system would enhance his or her job performance". On the other hand, "perceived ease of use" explains the user's perception of the amount of effort required to utilise the system or the extent to which a user believes that using a particular technology will be effortless. These two dimensions are still considered of striking importance for assessing *user* acceptance of technology, and we leveraged them in our model SAT.

2.2.3 Unified Theory of Acceptance and Use of Technology (UTAUT)

An evolution of the TAM model is the UTAUT model, namely "Unified Theory of Acceptance and Use of Technology". The UTAUT model enhances the dimensions of TAM, including variables that do not pertain to the actual use of the technology but, for example, on the social context where the technology and user interact.

UTAUT has four key constructs (i.e., performance expectancy, effort expectancy, social influence, and facilitating conditions) (see Figure 6) that influence behavioural intention to use a technology and/or technology use.

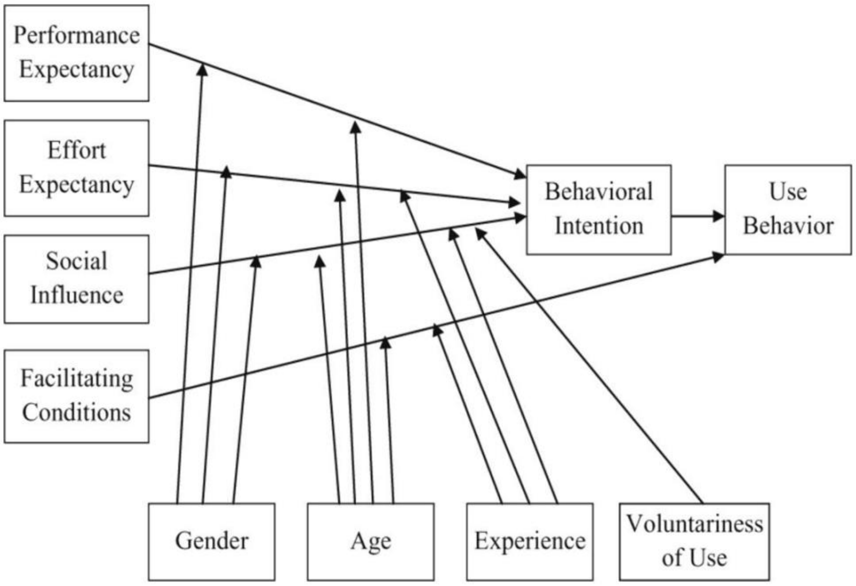


Figure 6: Unified Theory of Acceptance and Use of Technology (UTAUT)

It is clear that UTAUT is able to take into consideration multiple dimensions of the experience of technologies, and therefore assess the acceptance in a more holistic way, in respect to TAM. On the other hand, it is still strictly focused on user acceptance and highly correlated with the work environment. Moreover, it considers the relation between individuals-technologies-society as a one-way relation, without considering the socio-technical loop that entails these systems.

2.2.4 Technology, Organization, and Environment Framework (TOE)

This approach developed by (Tornatzky & Fleischer, 1990) is highly interesting since it takes into consideration the organisation and environmental features that enhance (or reduce) the adoption of a technology by a company. It offers a different point of view in respect to TAM or UTAUT, that are focused on users. In summary, TOE framework (see Figure 7) is focused technology (availability and characteristics), organisation (formal and informal linking structures, communication processes, size and slack), and environment (industry characteristics and market structure, technology support infrastructure and government regulation).

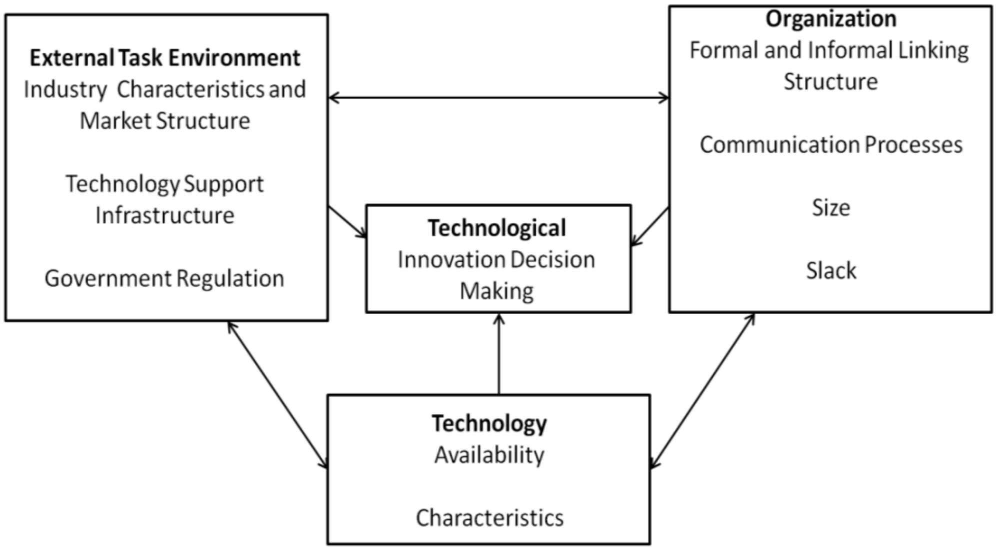

External Task Environment
Industry Characteristics and Market Structure
Technology Support Infrastructure
Government Regulation
Organization
Formal and Informal Linking Structure
Communication Processes
Size
Slack
Technological
Innovation Decision Making
Technology
Availability
Characteristics

Figure 7: Technology, Organization, and Environment Framework (TOE)

2.2.5 Diffusion of Innovation (DOI) Theory

The DOI theory (Rogers, 1995) develops a framework in order to understand which are the drivers that allow a technology to diffuse rapidly.

Diffusion of innovations (see Figure 8) is a theory that seeks to explain how, why, and at what rate new ideas and technology spread through cultures. Diffusion is the process in which an innovation is communicated through certain channels over time among members of a social system. It is a special type of communication in that the messages are concerned with new ideas. The four main elements in the diffusion of innovations are the innovation, communication channels, time and the social system. Diffusion occurs progressively within one market (a system of users) when information and opinions about a new technology are shared among potential users through communication channels.

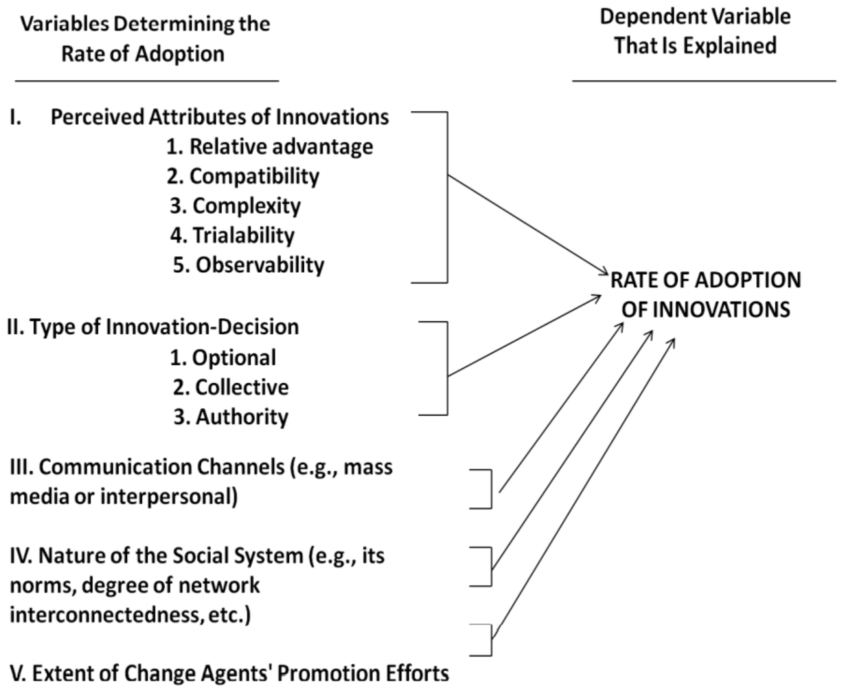

Variables Determining the Rate of Adoption
Dependent Variable That Is Explained
I. Perceived Attributes of Innovations
1. Relative advantage
2. Compatibility
3. Complexity
4. Trialability
5. Observability
II. Type of Innovation-Decision
1. Optional
2. Collective
3. Authority
III. Communication Channels (e.g., mass media or interpersonal)
IV. Nature of the Social System (e.g., its norms, degree of network interconnectedness, etc.)
V. Extent of Change Agents' Promotion Efforts
RATE OF ADOPTION OF INNOVATIONS

Figure 8: Diffusion of Innovations (DOI)

2.3 SAT methodology

The previous subsection 2.1 described the main and most adopted theoretical frameworks for human factors evaluation. They are all presenting some critical aspects, i.e., the lack of the organisational dimension of information sharing in the Kolini's conceptual framework, and the technical-peculiarity independent of the subjective perception in the Prat's model.

Instead, subsection 0 illustrated theories and models on technology acceptance, all with specific peculiarities: TAM focuses specifically on the user acceptance; UTAUT adopts TAM model specifying the social variables affecting technology acceptance; TOE is concerned mostly with the organisational and environmental aspects of technology adoption; DOI focuses on understanding the diffusion of technologies.

While all of them present some limitations that make them not suitable for the IRIS purpose of defining human factors for the assessment of its technology, the **Social Acceptance of Technology (SAT)** methodology (Occhipinti & et al., 2022), conceived by partner CEL as a result of its research experience (Briguglio & et al., 2021), goes beyond these models.

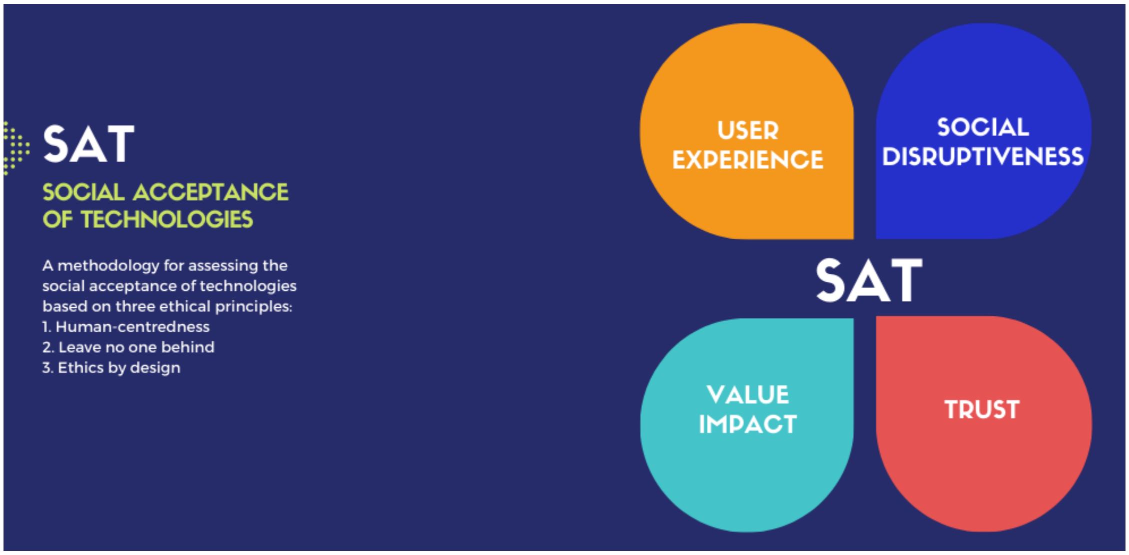
Social acceptance, as described and analysed in the previous sections, is a complex phenomenon that entails different dimensions: social, psychological, technology design, values of stakeholders, economic considerations etc. In this context and compared with the other theoretical models currently used in technology evaluation, the **SAT allows the evaluator to investigate and assess not only the correlation between behaviour**

and intention, but also how social influences acts in determining the individual's judgement of technology, and thus its acceptance. For a detailed description of the adopted items, see Table 4.

At the same time, the SAT methodology does not take a disembodied point of reference. It does not intend to replace the priority of the individual with the priority of the social. Instead, it intends to understand the relationship between these determinations from an understanding of the relationship with the technology at both individual and societal level.

> Therefore, **the SAT follows the lines of research of the TAM and UTAUT models in the assessment of technology acceptance, combining and enhancing them to take human factors into account in their socio-systemic and political aspects, with the aim of comprehensively assessing these dimensions.**

Therefore, while UTAUT and TAM are limited almost exclusively to the evaluation of the sum of individual users' experiences, SAT addresses the socio-technical and systemic nature of acceptance through **four conceptual constructs – called "bubbles" – that identify the four fundamental areas of evaluation** the method is based on.



A methodology for assessing the social acceptance of technologies based on three ethical principles:
1. Human-centredness
2. Leave no one behind
3. Ethics by design

Figure 9: Conceptual constructs of SAT methodology

The four bubbles (see Figure 9), are hereafter briefly described:

- **User Experience**: it aims to evaluate the user perceptions and expectations based on narrative and/or usage experience.
- **Value Impact:** it evaluates the extent to which the technology concerned complies with shared social values.
- **Perceived Trustworthiness:** it evaluates the extent to which the technological tool is considered reliable according to the individual user and to society.
- **Social Disruptiveness**: it is aimed at measuring, evaluating, and predicting the combination of three factors: the expected spread of a technology; how much it

will lead to a significant change from the point of view of production processes; how much it will impact on society as a socio-technical system.

Moreover, with the aim of being customised in the specific context, **the model itself is modular and scalable**, in the sense that while the SAT methodology integrates multiple methodologies that theoretically can assess the social acceptability of any technology, it can be applied picking just the necessary bubbles up, in case some bubbles might be out of scope or not very relevant for assessing a given technology.

3 IRIS SOCIAL ACCEPTANCE OF TECHNOLOGY

As depicted in section 2.3, the SAT methodology is a modular and scalable methodology that can be tailored to the specific technology and context.

The following subsections of this document will be devoted to the description of the customisation of the SAT methodology (subsection 2.3) for the IRIS specific purpose. The basic idea is that the technology is part of a socio-technical system that includes social value and cultural elements that determine its use, understanding, usefulness, and acceptability. From such an idea, the resulting methodology **called IRIS Social Acceptance Technology methodology (IRIS SAT)** will comprehend both the human factors (see Table 3) to be used for the assessment, as well as the assessment process that will be carried out in the IRIS project.

3.1 IRIS SAT methodology definition

The first steps for the customisation of the SAT, allowing to create a specific IRIS instance, consist on the identification of ***who*** are the stakeholders that will account for a successful co-design of IoT and AI enabled IRIS Virtual Cyber Range Platform, ***what*** will be assessed and ***how***, as follows:

- ***WHO***: security practitioners will be asked to assess the IRIS technology; security services providers and decision makers will be informed about assessment outcomes.
- ***WHAT***: expectations/perception of practitioners interacting with the IRIS Platform.
- ***HOW***: through the assessment of the four SAT bubbles, representing the IRIS human factor areas:

#	Human factor area	Description in the IRIS context
HFA1	**User Experience**	As described by (Friedli & Schuh, 2012), users are often neither involved nor consulted in the design process of new technologies that will then shape and modify the working environment. However, the relevant literature indicates that technologies with low user acceptance result in lower job satisfaction (Mariani & et al., 2013) and ultimately lead to under-performance (Devaraj & Kohli, 2003). In order to avoid the above-described risks, this area will take into account the user experience of the security practitioners.
HFA2	**Value Impact**	It will assess to what extent the IRIS technology and the organisation (e.g., the institution) that will adopt it adhere to shared social values, from the practitioner point of view.
HFA3	**Perceived Trustworthiness**	It will evaluate how reliable the technology in question is, in terms of transparency, certainty, risks and institutional trustworthiness.

#	Human factor area	Description in the IRIS context
HFA4	**Social Disruptiveness**	It will determine the impact of the IRIS technology in terms of increasing/decreasing of the practitioner’s security feeling in managing cyberthreats.

Table 3: IRIS Human Factor Areas

3.2 IRIS Human factors

The hereafter Table 4 will detail the selected human factors for each selected area, describing the corresponding barriers that hinder acceptance. Sources in literature, as well as among the ethics requirements on AI mechanisms included in D2.3 and reported in ANNEX I: ETHICS REQUIREMENTS ON TRUSTHWORTY AI (i.e., ECx sources) will be also identified.

#	Human Factor	Description and Barriers to overcome
		HFA1 User Experience
HFA1.1	**Perceived Usefulness (PU)**	How much practitioners perceive the technology as useful to its own field of work. **Barrier hindering acceptance**: Practitioners does not perceive the technology as useful. **Sources**: (Davis, 1989), (Venkatesh & Davis, 2000).
HFA1.2	**Perceived Ease of Use (PEU)**	How much practitioners perceive the technology as intuitive and easy to use. **Barrier hindering acceptance**: Practitioners perceive the technology as being too difficult to understand and make use of. **Sources**: (Davis, 1989), (Venkatesh & Davis, 2000).
HFA1.3	**Likability (LK)**	How much practitioners rate the technology on aspects such as enjoyable, entertaining, fun, appealing, interesting as well as overall like/dislike. **Barrier hindering acceptance**: Practitioners dislike the technology. **Sources**: (Lee & et al., 2011)

#	Human Factor	Description and Barriers to overcome
HFA1.4	**Reliability (RL)**	How much practitioners perceive that the technology does what it is supposed to do, performing according to its specifications. It is strongly related to PU, being however a specification on a more technical aspect of technology acceptance. **Barrier hindering acceptance**: Practitioners do not perceive the technology performing as expected. **Sources**: EC2 – Technical Robustness and Safety
		HFA2 Value Impact
HFA2.1	**Perceived Behaviour Control (PBC) - Human in the loop (HiL)**	The PBC expresses how much the user feels in control of the technology. Predicting both the behavioural intention (the propensity to adopt) and the adoption itself, it is used in the theory of planned behaviour to understand how the sense of self-efficacy of the users affects technology acceptance. In the IRIS' case, it is strictly related to the Human in the loop: the perception, from the practitioners' side, of being active participants of the process of the technology, and not mere spectators. It is one of the most demanded requirements in terms of the ethics of implementing AI systems. **Barrier hindering acceptance**: Practitioners do not perceive themselves as an active and participating user of the technology. **Sources**: (Ajzen, 1998), EC1 – Human agency and oversight
HFA2.2	**Capacity enabling (CE)**	How much practitioners perceive their capacities to be augmented by the use of the technology. It is a further elaboration on PBC and a refinement of PU, and it captures a more specific aspect of social acceptance, namely, the capacity to augment human capacities. **Barrier hindering acceptance**: Practitioners do not perceive their abilities increased by the technology. **Source**: EC1 – Human agency and oversight
		HFA3 Perceived Trustworthiness

#	Human Factor	Description and Barriers to overcome
HFA3.1	**Transparency (TR)**	How much the technology is believed to be understandable and its underling decision mechanisms not repudiable (in the case of AI, how much the black-box effect of the trained algorithm will be avoided). **Barrier hindering acceptance**: Practitioners do not perceive the technology to be understandable and accountable. **Sources**: EC4 - Transparency
HFA3.2	**User Perceived Certainty (SC)**	How much practitioners can foresee how the technology behave, what is its response and its impact on their work. This factor represents an elaboration of the "Observability" item in (Sahin, 2006) **Barrier hindering acceptance**: Practitioners are not able to predict the outcomes of the technology. **Sources**: (Sahin, 2006)
HFA3.3	**Perceived Risks (PR)**	How much practitioners perceive that the technology (a malfunctioning, or a malicious intervention on it) might represent a threat once adopted. **Barrier hindering acceptance**: Practitioners believe the technology could harm someone/something. **Sources**: (Covello, 1983)
HFA3.4	**Institutional Trustworthiness (ITW)**	How much practitioners perceive the social and political environment in which they operate to be trustworthy in terms of cybersecurity. **Barrier hindering acceptance**: Practitioners do not trust the organisations operating and controlling the technology. **Sources**: EC7 - Accountability
		HFA4 Social Disruptiveness
HFA4.1	**Expected systemic change (ESC)**	How much practitioners believe the technology might deeply change the way cyberthreats are handled. **Barrier hindering acceptance**: Practitioners do not believe the technology will make significant changes in a long-term period. **Sources**: (Maturana & Varela, 1991)

Table 4: IRIS Human Factors

3.3 IRIS Assessment tools and techniques

The IRIS adopted methodology will apply both **Qualitative** and **Quantitative** methods, using two main empirical study techniques (i.e., questionnaires and focus groups). Details will be defined in the next project period and reported in D2.7, due at month 24.

3.3.1 Qualitative assessment

From a qualitative point of view, the IRIS SAT methodology will make use of two differentiated qualitative methods, including:

- **Semi-structured interview**: it is based on the formulation of questions to adequately explore the domains of interest of a given research. Structured interviews take place when the researcher asks predetermined questions in a closed manner on certain items that are to be investigated, and which have been previously outlined in the theoretical research phase (literature analysis).
- **Focus Group**: it is a specific form of group interview, in which the interviewer coordinates a limited number of people by stimulating their interaction, communication and dialogue. It has a number of elements such as the centrality of the group as a source of information, the interaction of the subjects, the focus on a specific topic.

3.3.2 Quantitative assessment

The IRIS SAT methodology will also make use of quantitative methods, differentiated according to the human factor dimensions, that will be used to statistically investigate the responses relevant to a given field.

The questionnaires are designed with the following rationale: they will be based on a *likert scale* and will have three response options, two aimed at investigating a certain content, and the other as a response check - sometimes in a negative form - to ascertain the respondent's attention, also allowing to weigh the answers given to the previous two questions. Stakeholders will be asked to answer all the questions.

3.3.3 IRIS Questionnaire

As explained in the previous section 3.1, the IRIS SAT methodology will make use of 11 human factors divided into 4 different areas, with the aim of investigating how to overcome the identified related barriers. A questionnaire will be submitted to the security stakeholders. It contains 3 questions for each identified human factor, in a way that the third question will help to avoid the response set effect (Hasshim & et al., 2018). The questionnaire is listed hereafter. It should be noticed that it is a first proposal and will be refined in the next project phases and reported in D2.7 at month 24.

#	Human Factor	Question
		HFA1 User Experience
HFA1.1	**Perceived Usefulness (PU)**	• I think this technology can be useful in my working sphere • I think this technology can be useful in my daily life • I think this technology may be of little use to me
HFA1.2	**Perceived Ease of Use (PEU)**	• I find this technology intuitive • I think I would quickly learn how to use this technology • I think it's hard to understand how this technology works
HFA1.3	**Likability (LK)**	• I would like to adopt this technology • I find this technology smart and nice • I find that this technology is not pleasant
HFA1.4	**Reliability (RL)**	• I feel that this technology only works as it is supposed to do • I feel I can rely on the functioning of this technology without worries • This technology gives me the feeling of working randomly
		HFA2 Value Impact
HFA2.1	**Perceived Behaviour Control (PBC) - Human in the loop (HiL)**	• I feel to be fully in control of this technology • I feel comfortable using this technology • I feel that the effects of this technology are beyond my control
HFA2.2	**Capacity enabling (CE)**	• I believe that this technology gives me a sense of ability and efficacy. • I believe that this technology makes easier for me to reach my goals • I believe that that this technology does not improve my abilities
		HFA3 Perceived Trustworthiness
HFA3.1	**Transparency (TR)**	• I think that the behaviours of this technology are comprehensible for me • I think that this technology is well documented and explained • I feel that the functioning of this technology is obscure to me
HFA3.2	**User Perceived Certainty (SC)**	• I know what is going on when the technology is working • I feel I can predict the effects of this technology to me and to the outer environment • I know how to modify the working of this technology to make it follows my wills
HFA3.3	**Perceived Risks (PR)**	• I think this technology is risky for me • I think this technology can harm someone/something • I think that the impact of the risk associated to this technology is relevant to me

#	Human Factor	Question
HFA3.4	**Institutional Trustworthiness (ITW)**	• I believe that the regulator that is in charge of controlling this technology is worthy of my trust • I think that the manufacturer of this technology is considerably trustworthy • In general, I believe that I cannot trust the party involved in making and controlling this technology
HFA4 Social Disruptiveness		
HFA4.1	**Expected systemic change (ESC)**	• I believe that this technology will deeply change the way cyberthreats are handled • I think that I believe that this technology will start a process on cyberthreats handling that cannot be stopped • I believe that this technology will not have impacts in a long-term period

Table 5: IRIS Questionnaire

3.4 IRIS Assessment process

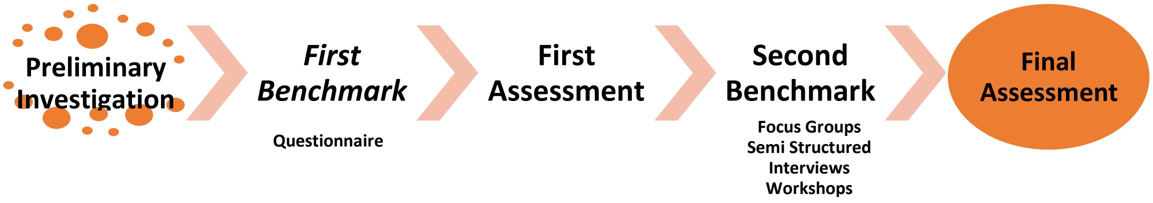


Figure 10: IRIS SAT Process

In order to ensure that the technology is developed ensuring a co-creation process for an inclusive and open technology development, the methodology process will be implemented in two iterations, with continuous feedback, as follows:

1. **First Iteration**
   1.1. **Preliminary investigation** and selection of human factors performed in the current document.
   1.2. **First benchmark** via questionnaire to investigate how the technology is perceived, or expected, from the practitioners (1st practitioner investigation).
   1.3. **First elaboration** to analyse the questionnaire's results and first round of recommendations (1st internal meeting).
2. **Second Iteration**
   2.1. **Second benchmark** via Focus Group through the discussion of relevant dimensions for the interaction with technology.
   2.2. **Final elaboration and recommendations** to determine how barriers might be overcome by the project, to make final recommendations.

The first iteration will provide the IRIS developer partners with the data from the first surveys and questionnaire results so that they can refine and improve the technology and its implementation.

Subsequently, the second iteration of the methodology will consist of an overall assessment of the development of the technology and its implementation in the project, in order to evaluate further aspects during project development and the progressive definition of the technology and its use.

All the collected data will be finally analysed to draw up an overall assessment of the project's progress, and presented to a team of ethics experts (e.g., the project Ethics Board) for their validation and any eventual further thoughts or concerns.

4 CONCLUSIONS

This deliverable proposes an open, flexible, and scalable methodology. It attempts to understand and evaluate human, social, and cultural factors with the aim of understanding their role and relevance in the implementation of the IRIS project technology.

Human factors and their evaluation are indeed crucial in relation to AI development as well as in relation to cybersecurity and information sharing.

In fact, a careful consideration of their role allows us to go beyond the existing literature on cybersecurity and the evaluation of technology, by including a dimension (that relating to human factors) understood in a way that can account for both the individual and social dimensions as well as the political and cultural ones.

The document shows how, starting from analyses of the existing literature, especially with reference to the TAM and UTAUT models and the definition of human factors (TOE framework) in information sharing, the proposed methodology is meant to go beyond the state of the art.

In fact, it includes aspects related to individual and psychological perceptions of the user and technology acceptance, as well as social and political aspects, with the aim of comprehensively assessing them. All of these aspects are definitely intertwined and for this reason the proposed methodology includes them for performing a holistic assessment of the technology.

The methodology will be applied in two iterations with continuous feedback from the stakeholders for the analysis of technology design and implementation. Results will inform project partners developers about the gaps and the areas of improvement.

In addition, it will make use of methodological tools for the quantitative and qualitative assessment.

Therefore, the Deliverable achieves the following objectives:

1. Provide an overview of the state of the art of the current human factors and technology acceptance assessment panorama, highlighting potentialities and limitations;
2. Goes beyond the existing state of the art by including social, political and value aspects, as well as the psychological ones, in the methodology for assessing human factors and technology acceptance;
3. Develops an open, flexible and scalable methodology (including 11 human factors classified in 4 categories, relying to the four bubbles in the SAT methodology) that will be implemented during the course of the project, in close contact with security practitioners, security developers and other relevant stakeholders (ethicists, institutional players, policy makers);
4. List and motivate the qualitative and quantitative methodological techniques that will be used (including a questionnaire template for each human factor), describing the iterations with continuous feedback that will be put in place to ensure co-creation and stakeholder engagement.

In conclusion, the proposed methodology, which is based on the experience and lessons learnt of CEL in other research activities, already validated in its first stage by the project Ethics Board, will be developed and refined in the course of project research, in continuous coordination with task T2.6, which will see its direct application.

5 REFERENCES

AI HLEG. (2019). *Ethics Guidelines for Trustworthy AI.* (AI), High-Level Expert Group on Artificial Intelligence.

Ajzen, I. (1998). Models of human social behavior and their application to health psychology. *Psychology & Health*, 735–739.

Boyce, M., & et al. (2011). Human Performance in Cybersecurity: A Research Agenda. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, (pp. 1115-1119).

Briguglio, L., & et al. (2021). Business Value and Social Acceptance for the Validation of 5G Technology. *2021 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)* (pp. 132-137). Athens: IEEE.

Corradini, I. (2020). *Building a Cybersecurity Culture in Organizations, How to Bridge the Gap Between People and Digital Technology.* Springer.

Covello, V. (1983). The Perception of Technological Risks: A Literature Review. *Technological Forecasting and Social Change*, 285-297.

Davis, F. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly: Management Information Systems*, 319-339.

Devaraj, S., & Kohli, R. (2003). Performance impacts of information technology: Is actual usage the missing link? . *Management Science*, 273–289.

Fishbein, M., & Ajzen, I. (1975). *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research.* Addison-Wesley.

Friedli, T., & Schuh, G. (2012). *Wettbewerbsfähigkeit der Produktion an Hochlohnstandorten [Competitiveness of production at high-wage locations].* Berlin, Germany: Springer.

Hasshim, N., & et al. (2018). Trial type mixing substantially reduces the response set effect in the Stroop task. *Acta Psychol (Amst)*, 43-53.

Jeong, J., & et al. (2019). Towards an Improved Understanding of Human Factors in Cybersecurity," 2019. *Proceedings of the IEEE 5th International Conference on Collaboration and Internet Computing (CIC)*, (pp. 338-345).

Kim, Y., & Crowston, K. (2011). Technology Adoption and Use Theory Review for Studying Scientists’ Continued Use of Cyber-Infrastructure. *Proceedings of the American Society for Information Science and Technology*, 1-10.

Kolini, F. (2017). Two Heads are Better than One: A Theoretical Model for Cybersecurity Intelligence Sharing (CIS) between Organisations. *The Sixth Asian Conference on Information Systems (ACIS2017) Proceedings*, (p. 88).

Lee, S., & et al. (2011). Effects of Appearance and Functions on Likability and Perceived Occupational Suitability of Robots. *Journal of Cognitive Engineering and Decision Making*, 232–250.

Mariani, M., & et al. (2013). Training opportunities, technology acceptance and job satisfaction: A study of Italian organizations. *Journal of Workplace Learning*, 455–475.

Maturana, H., & Varela, F. (1991). *Autopoiesis and cognition: The realization of the living.* Springer Science & Business Media.

Occhipinti, C., & et al. (2022). SAT: a methodology to assess the social acceptance of innovative AI-based technologies. *Journal of Information, Communication and Ethics in Society*.

Oliveira, T., & Martins, M. (2011). *Literature Review of Information Technology Adoption Models at Firm Level.*

Prat, N., & et al. (2014). *Artifact evaluation in information systems design-science research–a holistic view.*

Ramayah, T., & Jantan, M. (2004). Technology acceptance: an individual perspective. Current and future research in Malaysia. *Review of Business Research*, 103-111.

Rogers, E. (1995). *Diffusion of Innovations.* New York: The Free Press.

Ropohl, G. (1999). Philosophy of socio-technical systems. *Society for Philosophy and Technology Quarterly Electronic Journal*, 186-194.

Sahin, I. (2006). Detailed review of Rogers' diffusion of innovations theory and educational technology-related studies based on Rogers' theory. *Turkish Online Journal of Educational Technology-TOJET*, 14-23.

Tornatzky, L., & Fleischer, M. (1990). *The Processes of Technological Innovation.* Lexington: Lexington Books.

Venkatesh, V., & Davis, F. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science*, 186–204.

Venkatesh, V., & et al. (2016). Unified Theory of Acceptance and Use of Technology: A Synthesis and the Road Ahead. *Journal of Association for Information Systems*, 328–376.

ANNEX I: ETHICS REQUIREMENTS ON TRUSTHWORTY AI

The following table lists the constraints from the *Ethics Guidelines for trustworthy AI* (AI HLEG, 2019) depicted in D2.3.

Req #ID	Trustworthy AI constraint	Potential Risk	IRIS Requirement
ER1	EC1 - Human agency and oversight	- The subject is unable to make autonomous and informed choices - Subject's dignity as an agency person is violated	Human in the loop and Human in command mechanisms shall be implemented
ER2	EC2 - Technical Robustness and safety	- The system could be used by malicious actors - In case of damage, if there is no fallback plan, the damage may extend to things, people, environment - The system may not provide correct and accurate indications and information - If the system does not have a high rate of reproducibility, it may be unpredictable	- Non-repudiation mechanisms shall be implemented - An accurate test plan to be reproduced over time to ensure the efficiency and proper functioning of the system shall be prepared, so that the degree of accuracy and reproducibility can be checked and verified - System stakeholders shall be adequately informed e.g., throw adequate informative material
ER3	EC3- Privacy and data governance	- Risks are highlighted in the section on data protection and governance (see 0)	The actions required to mitigate these risks are highlighted in the section on GDPR requirements (see 0)

Req #ID	Trustworthy AI constraint	Potential Risk	IRIS Requirement
ER4	EC4-Transparency	-The system is difficult to explain and understand	As the information processed by the IRIS platform is strictly confidential and relevant to security issues, processes and system behaviour (both technical and decision making) shall be carefully documented and tracked to ensure transparency
ER5	EC5- Diversity, non-discrimination and fairness	- The presence of discriminatory bias leads to actions that may marginalize and discriminate against certain groups or categories of people - Non-universal design may exclude certain categories of people (e.g., people with disabilities) - If stakeholders are not involved, the system may be developed in an undemocratic way	- Decision-making processes shall not be made based on discriminatory bias. A group of external experts shall be consulted to make assessments and analyses of possible discriminatory biases - The platform interface and functionalities shall be universally accessible to all human beings, respecting their diversity - Co-design involving all relevant stakeholders' categories shall be ensured

Req #ID	Trustworthy AI constraint	Potential Risk	IRIS Requirement
ER6	ER6- Societal and environmental well-being	- The system might harm not only people, but also other sentient beings, the environment and the society as a whole - If adequate measures are not taken, the impact of the AI system on the mental and physical well-being of people and the community may not be properly assessed	The system shall be sustainable from an environmental and energetic point of view, being compliant with the Do Not Significant Harm (DNSH)[1] principle
ER7	EC7- Accountability	- Without appropriate auditability and redress measures, the system might be considered untrustworthy - It might be difficult to trace processes	- A lead manager who is responsible for the AI system who can account for the consequences of actions taken shall be identified and communicated to the stakeholders - A tracking mechanism shall be implemented to log accesses and actions carried out by using the system

Table 6: EU Guidelines for trustworthy AI constraints

[1] https://ec.europa.eu/info/sites/default/files/2021_02_18_epc_do_not_significant_harm_-technical_guidance_by_the_commission.pdf